From Words to Behaviour via Semantic Networks

Armand S. Rotaru (ucjt401@ucl.ac.uk)  Gabriella Vigliocco (g.vigliocco@ucl.ac.uk)
Faculty of Brain Sciences, University College London,
WC1H 0DS, London, United Kingdom

Stefan L. Frank (s.frank@let.ru.nl)
Centre for Language Studies, Radboud University,
6525 HT Nijmegen, the Netherlands

Abstract
The contents and structure of semantic networks have been the focus of much recent research, with major advances in the development of distributional models. In parallel, connectionist modeling has extended our knowledge of the processes engaged in semantic activation. However, these two lines of investigation have rarely brought together. Here, starting from a standard textual model of semantics, we allow activation to spread throughout its associated semantic network, as dictated by the patterns of semantic similarity between words. We find that the activation profile of the network, measured at various time points, can successfully account for response times in the lexical decision task, as well as for subjective concreteness and imageability ratings.

Keywords: computational modelling; semantic networks; text corpora; lexical decision; concreteness; imageability

Introduction
In the last 15 years, a great deal of effort was invested in collecting extensive behavioural norms, for lexical semantic tasks such as free association (Nelson, McEvoy, & Schreiber, 2004), similarity judgement (Bruni, Tran, & Baroni, 2014; Silberer & Lapata, 2014), feature generation (McRae, Cree, Seidenberg, & McNorgan, 2005; Recchia & Jones, 2012; Vinson & Vigliocco, 2008). In addition, large norms have been obtained for tasks that rely primarily on orthographic and phonological processing, but also include a semantic component, such as lexical decision (Balota et al., 2007; Keuleers, Lacey, Rastle, & Brysbaert, 2012).

This wealth of data has allowed researchers to start exploring the ties that link language to perception and action, in a more methodical and in-depth manner than was previously possible. At a general level, especially within the fields of computational linguistics and natural language processing, representational similarity analysis has been employed in order to study verbal and visual semantic representations across domains of knowledge (Kriegeskorte, Mur, & Bandettini, 2008; for a recent review, see Kriegeskorte & Kievit, 2013). This approach is inspired by several embodied theories of cognition in which the semantic system is considered to rely on integrated modal (especially visual) and amodal representations (Barsalou, Santos, Simmons, & Wilson, 2008; Louwerse, 2007; Vigliocco, Meteyard, Andrews, & Kousta, 2009). The research following said approach has shown that unimodal (i.e., verbal or visual), but especially multimodal (i.e., verbal-visual) distributional models (for a detailed review, see Bruni, Tran, & Baroni, 2014) can provide a good account of human task performance in a number of semantic tasks. Such studies demonstrated that integrating information from two modalities provides a better account of behavioural data than that offered by the individual modalities, across a wide range of models and integration methods, even for abstract concepts, such as peace and freedom (Bruni, Tran, & Baroni, 2014; Hill & Korhonen, 2014; Hill, Reichart & Korhonen, 2014). The results are consistent with those of previous studies (Andrews, Vigliocco, & Vinson, 2009; Louwerse, 2011; Maki & Buchanan, 2008; Riordan & Jones, 2011; Sadeghi, McClelland, & Hoffman, 2015; Steyvers, 2010), indicating that language and perception can be seen as highly redundant, yet complementary, sources of semantic information.

Differences in the reliance upon one or the other modalities, as well as in degree and strength of association to other concepts, have been argued to underscore difference across domains of knowledge. For example, representational richness has been argued to underlie the distinction between concrete and abstract concepts, whereby concrete concepts are richer than abstract ones when it comes to perceptual and motor elements, but poorer with respect to introspective and linguistic elements (see Gee, Nelson, & Krawczyk, 1999; Hill, Korhonen, & Bentz, 2014; Pecher, Boot, & Van Dantziz, 2011; Vigliocco et al., 2009; Wiemer-Hastings & Xu, 2005). A large number of studies have used comprehensive behavioural norms and subjective ratings to evaluate the role of semantic richness, using different measures of richness such as number of features as well as contextual and semantic diversity, to name a few (for reviews, see Jones, Johns, & Recchia, 2012; Mirman & Magnuson, 2008; Yap, Pexman, Wellsby, Hargreaves, & Huff, 2012). Not surprisingly, the results paint a rather complex picture, where semantic richness effects are both task-general and task-specific, have both an early and a late impact on task behaviour (Hargreaves & Pexman, 2014), and either facilitate or hinder task performance (Mirman & Magnuson, 2008).

Here, we attempt to bring a fresh perspective in the study of how concepts (both concrete and abstract) are represented and, crucially, processed, by developing a computational model that accounts for previous findings by incorporating
both structural and dynamical elements. In particular, we explore the extent to which we can predict response times and accuracies in visual word recognition (i.e., lexical decision), as well as both concreteness and imageability ratings, starting from distributional models of semantics (Mandera, Keuleers, & Brysbaert, 2015; Westbury et al., 2013) supplemented by simple assumptions concerning the dynamic spreading of activation during processing.

**Method**

**Model**

We derive semantic richness measures of words from a probabilistic model of semantic processing, in the following manner: (a) pre-process the written part of the British National Corpus (Leech, Garside, & Bryant, 1994), by converting all the words to lowercase, eliminating punctuation marks and removing words whose absolute frequencies are less than 5; (b) construct 300-dimensional vector representations for the words in the BNC, by employing the Skipgram model (Mikolov, Chen, Corrado, & Dean, 2013); (c) compute a representational similarity matrix \( DM \) from said vectors, using vector cosine as a measure of similarity between the vectors (i.e., words); (d) set to zero all the negative values in \( DM \), as a means of reducing the amount of noise present (i.e., vector cosines which carry very little semantic information); (e) normalize the rows of the matrix, such that each row sums to one, and that any value \( DM(I,J) \) can be interpreted as the strength of the directional connection from word \( W_I \) to word \( W_J \); (f) consider the discrete Markov chain associated with \( DM \), which we denote as \( MARKOV(DM) \), and compute the state of \( MARKOV(DM) \) at steps \( K = 1 \) through \( K = 7 \), namely \( S_k(DM) \); (g) for each word \( W \) and each \( K \) between 1 and 7, count the number of close neighbours of \( W \) (\( numNeigh_1 \)). A word \( V \) is considered a close neighbour of \( W \) if \( P(V | S_k(DM)) > threshold_k \), where \( threshold_k \) are lower thresholds.

In the end, we are left with seven free parameters (i.e., \( threshold_1 \)) and seven semantic richness measures (i.e., \( numNeigh_{1-7} \)), as well as with a few fixed parameters for the underlying Skipgram model. Although our richness measures are all derived in a very similar manner, they have rather different interpretations, at least from a graph-theoretical perspective (Koschützki, Lehmann, Peeters, Richter, Tenfelde-Podehl, & Zlotowski, 2005). The meaning of each measure is briefly described in Table 1.

**Data Analysis**

We focus on four dependent measures: concreteness ratings (Brysbaert, Warriner, & Kuperman, 2014), imageability ratings (Gilhooly & Logie, 1980; Stadthagen-Gonzalez & Davis, 2006), and both accuracies and response times from a lexical decision task, for a subset of 2,328 words from Keuleers, Lacey, Rastle, and Brysbaert (2012).

We include the following baseline variables: (log) contextual diversity, (log) frequency (van Heuven, Mandera, Keuleers, & Brysbaert, 2014), familiarity (Gilhooly & Logie, 1980; Stadthagen-Gonzalez & Davis, 2006), age of acquisition (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012), (squared) hedonic valence (Warriner, Kuperman, & Brysbaert, 2013), number of letters, Coltheart’s N (i.e., the number of words that can be produced by substituting one letter of a given word for any other, such that the result is a valid word; Coltheart, Davelaar, Jonasson, & Besner, 1977), orthographic Levenshtein distance (OLD20; the average orthographic editing distance between a word and its twenty closest neighbours in the lexicon; Yarkoni, Balota, & Yap, 2008), and phonological Levenshtein distance (PLD20; the average phonological distance between a word and its twenty closest neighbours in the lexicon; Suárez, Tan, Yap, & Goh, 2011). In addition we include semantic diversity (Hoffman, Lambon Ralph, & Rogers, 2013) as a baseline measure. This latter has been argued to capture basic semantic differences across concepts as represented in distributional semantic networks. Our variables of interest are the seven measures of semantic richness (i.e., \( numNeigh_{1-7} \)).

We run two multiple linear regressions, one for the baseline variables, and one for the complete set of predictors (i.e., the baseline variables and our semantic richness measures). Since our richness measures are very strongly correlated with one another, we partial out any variance shared with other predictors, such that \( numNeigh_{R_k} = Res \)

---

1 We prefer the Skipgram model for two main reasons: firstly, it is nearly state-of-the-art when it comes to accounting for behavioral data (Baroni, Dinu, & Kruszewski, 2014; Mandera, Keuleers, & Brysbaert, 2015); secondly, several freely available, computationally efficient and well documented implementations of the model exist (https://code.google.com/p/word2vec/).

2 We use the Skipgram implementation provided by the gensim tool (Rehurek & Sojka, 2010), with the following parameter values: \( alpha = 0.025 \) (initial learning rate), \( window = 5 \) (radius of sliding window), \( sample = 0 \) (amount of downsampling), \( negative = 0 \) (amount of negative sampling), and \( iter = 1 \) (number of iterations over the entire corpus).
(numNeighK ~ Baseline + numNeighR1 + … + numNeighRK−1), for all values of K between 1 and 7. Therefore, our predictors consist of Baseline and numNeighR1,7, whereas our dependent variables are Log RT, Accuracy, Concreteness and Imageability.

We employ one half of the words for model tuning, and the other half for model evaluation. We derive the optimal values for our predictors by using a variant of the simplex method (Lagarias, Reeds, Wright, & Wright, 1998), with (negative) total amount of variance explained serving as the objective function. In order to avoid local minima, we run 100 iterations of the optimisation process, and keep only the best result.

**Results**

The results are displayed in Tables 2 and 3. Our semantic richness measures can account for a significant amount of variance in concreteness and imageability ratings, as well as in response times in the lexical decision task. However, they do not explain variance in lexical decision accuracy over and above the baseline measures (Table 2). Table 3 shows the regression weights for all predictors and dependent variables.

Table 2. Percentage of variance accounted for by two models: a baseline model, and a combined one, consisting of all the predictors in the baseline model plus the semantic richness measures numNeighR1 through numNeighR7 (all values are significant at .001 level, except for accuracy in the “combined – baseline” comparison)

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Baseline</th>
<th>Combined</th>
<th>Combined – baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log RT</td>
<td>47.93</td>
<td>49.80</td>
<td>1.87</td>
</tr>
<tr>
<td>Accuracy</td>
<td>27.09</td>
<td>27.81</td>
<td>0.72</td>
</tr>
<tr>
<td>Concreteness</td>
<td>35.40</td>
<td>58.59</td>
<td>23.19</td>
</tr>
<tr>
<td>Imageability</td>
<td>31.90</td>
<td>53.24</td>
<td>21.34</td>
</tr>
</tbody>
</table>

**Discussion and Conclusions**

We develop a model that takes into account both the structural properties of semantics networks, as well as their dynamic aspects, by considering the flow of semantic activation generated by the automatic processing of individual words. An important result of looking at both structure and dynamics is that it allows us to assess the effects of both direct and indirect, mediated semantic relations between words, rather than limiting our analysis to strong, direct semantic links. Our results suggest that the explanatory power of text-based semantic representations is currently being underestimated, as a consequence of not taking into consideration the additional information provided by spreading activation mechanisms. By ignoring these simple processes, the extra information they generate would have to be integrated into the representations by
design, which would lead to the conflation of representations and processes.

Based on the results presented in Table 2, it seems that our model is considerably more suitable for predicting concreteness and imageability ratings, than reaction time and accuracy in the word recognition task. We believe that this phenomenon might be due to differences between the requirements of the lexical decision task on the one hand, and those of the concreteness/imageability rating task, on the other. Since our model assumes that the string of letters received as input is already a word, it is not surprising that it fares rather poorly in predicting lexical decision response time and accuracy. In contrast, the rating task involves making a considerably more elaborate discrimination, one between concrete/imageable and abstract/non-imageable words, all of which are present in our model (Buchanan, Westbury, & Burgess, 2001).

Beyond the promising initial results, we believe that our model has a number of advantages, which recommend it as a potentially useful tool in the study of semantic processing. In our opinion, the main quality of our model is that it ties together a number of competing modelling approaches, and combines many of their strengths, while avoiding most of their limitations.

Firstly, our model has a pronounced connectionist and/or dynamical systems flavour to it (Anderson, 1983; for a review, see McClelland et al., 2010), whereby the dynamics of the model can be interpreted in terms of “spreading activation”. In this case, activation flows from an initial concept to its neighbours, then to the neighbours of its neighbours, and so on, until the system reaches a global “attractor” state (i.e., an eigenstate). However, unlike other existing models (Chen & Mirman, 2012; Hoffman & Woollams, 2015; Rogers & McClelland, 2004), it has a large number of nodes and feedforward/feedback/recurrent connections, making it slightly more realistic and comprehensive. As a result, it might provide better insight into the distinct contribution of structural and task related aspects of semantic behaviour. One potentially promising approach in this regard comes from network science and the theory of stochastic processes, two methodologies which have attracted an increasing amount of attention in cognitive science (De Deyne & Storms, 2008; Ferrer i Cancho & Solé, 2001; Gruenenfelder, Recchia, Rubin, & Jones, in press; Steyvers & Tenenbaum, 2005; Utsumi, 2015; for a general review of network-based analyses of cognition, see Baronchelli, Ferrer i Cancho, Pastor-Satorras, Chater, & Christiansen, 2013). Another possibility might be to use a respond-to-signal paradigm (Ratcliff, 2006; Hargreaves & Pexman, 2014), which would provide additional quantitative insights on the accumulation of task-specific and task-independent information during task performance (e.g., in the word naming or the lexical decision tasks).

Secondly, our model can be seen as a probabilistic one (Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010), such that at each step, the model makes use of its underlying Markov chain, namely MARKOV(DM), in order to perform multi-step inferences. In contrast to other probabilistic models, such as Topics (Griffiths, Steyvers, & Tenenbaum, 2007), our model is non-hierarchical and does not undergo any form of dimensionality reduction, which means that the inferences are easier to interpret and that less semantic information is lost. Said inferences allow us to assess the strength of both direct and indirect semantic relations between words (Steyvers, Shiffrin, & Nelson, 2004; Howard, Shankar, & Jagadisan, 2011), for instance by testing whether certain words and/or associations between words are crucial for successfully carrying out a semantic task. Moreover, we can also examine the manner in which semantic cues restrict and guide the inference process, as is the case in tasks such as semantic fluency (Hills, Jones, & Todd, 2012), continued free association (De Deyne & Storms, 2008), and extralist cued recall (Nelson, Kitto, Galea, McEvoy, & Bruza, 2013).

Finally, our model is relatively simple, from a structural point of view, and is completely transparent in terms of its parameters. Taken together, these features make our model easy to run, and facilitate comparisons across different subsets of participants, stimuli and tasks. Also, as a results of its simplicity, our current model is very much open to extensions, for instance in order to increase its neuropsychological plausibility.

References


Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known
English word lemmas. *Behavior Research Methods, 46*(3), 904-911.


McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: connectionist and
dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, 14(8), 348-356.